BeNeLux Bioinformatics Conference – Brussels, December 2013
Abstract ID: **000** Category: Abstract template

BI-CLUSTERING GENE EXPRESSION DATA UNDER CONSTRAINTS

Thanh Le Van[1,], Ana Carolina Fierro[2], Tias Guns[1], Matthijs van Leeuwen[1], Siegfried Nijssen[1], Luc De Raedt[1], and Kathleen Marchal[2,3].*
*Depts. of Computer Sciences[1], and Microbial and Molecular Systems[2], KU Leuven ; Dept. of Plant Biotechnology and Bioinformatics[3], Ghent University. *thanh.levan@cs.kuleuven.be*

This paper presents a constraint-based approach to mining bi-clusters in gene expression data. Instead of designing an algorithm for each specific task, we propose to use constraint programming to turn the mining problem into a constraint satisfaction and/or optimisation problem. We demonstrate this promising approach on two cases. The first is to mine a single constant-row bi-cluster under noise constraints. The second is to mine a set of generic noisy constant-row bi-clusters under structure constraints, which is called a staircase pattern.

INTRODUCTION

In gene expression analysis, we are given a data matrix in which rows correspond to genes, columns correspond to conditions and data shows expression values of genes in conditions. A bi-clustering algorithm typically finds a subset of genes that shows an approximately constant value for a subset of conditions. The submatrix formed by the selected subset of genes and conditions is called a bi-cluster. Bi-clusters are interesting as the relationship between conditions and genes provides insight in the correlation of genes and can be used for finding perturbed biological processes or predicting gene regulation networks. The challenge that we study is to develop a generic and extendible approach to take into account requirements of a good bi-cluster. Some desirable properties include: rows need to be approximately constant; bi-clusters do not have much noise; constraints can be easily added or removed when we want to perform integrative data analysis.

METHODS

Different from earlier approaches, we propose a more general way to formalize and solve the problem using constraints. The prominent contribution lies in the fact that the entire model, which consists of an objective function and a set of constraints, is specified in a declarative programming language and solved using existing techniques supported by the language. In practice, we chose the constraint programming (CP) paradigm for modelling and solving. Working in this way, we have a number of advantages. First, it is a declarative approach. We can exploit built-in solving capabilities implemented by CP solvers, for instance constraint propagation and search strategies, to avoid re-inventing the wheel for common tasks and have more time to focus on modelling. The program we build will be easier to maintain or extend. Second, it is not hard to extend the model to other settings, for example, detected bi-clusters should have consistent patterns in another data matrix.

We select two settings for the mining task to present the proposed methodology.

In the first setting, we demonstrate how to compile the problem of mining a single fault-tolerant constant-row bi-cluster that covers the largest part of the data into a constraint optimisation problem. We also show how to use large neighbourhood search, a type of local search, to approximate the optimal solution.

In the second setting, we illustrate the extensibility of the framework to pattern set mining under structure constraints. More precisely, we want to find a set of fault-tolerant constant-row bi-clusters which resembles a staircase [1]. As the quality of the staircase depends on the user-defined noise thresholds, we propose a two-phased mining approach. First, we generate staircase candidates by solving a number of constraint satisfaction problems. Then, we use the Minimum Description Length (MDL) principle to select the best one. According to the MDL principle, the best model is the one that compresses the data best. In this case, a model is a staircase.

RESULTS & DISCUSSION

We experimented with a number of synthetic data sets of 1000 rows and 120 columns with varying noise levels and a number of staircase steps. Figure 1 shows that our model can recover most of the staircase and the MDL scores help us to select the best candidate.

In real data sets, we encountered the scalability problem of solvers. Besides that, detected staircases do not often have discernible steps (high noise outside). In the future, we plan to integrate with more data sets to increase the quality of the bi-clusters.

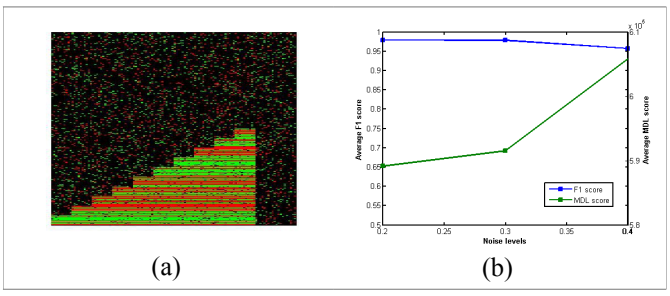

***FIGURE 1**. a) A synthetic staircase. b) Relating F1-scores to MDL-scores*

REFERENCES

1 Le Van, T., Fierro Gutiérrez, A., Guns, T., van Leeuwen, M., Nijssen, S., De Raedt, L., Marchal, K. (2012). Mining local staircase patterns in noisy data. *12th IEEE International Conference on Data Mining Workshops*. International workshop on Co-Clustering and Applications (CoClus'12) in conjunction with IEEE ICDM 2012.